Corela
Cognition, représentation, langage

5-1 | 2007
Vol. 5, n° 1

NdeN et acquisition d'informations lexicales à partir du Trésor de la Langue Française Informatisé

Laurence Kister et Evelyne Jacquey

Édition électronique
URL : http://journals.openedition.org/corela/332
DOI : 10.4000/corela.332
ISSN : 1638-573X

Éditeur
Cercle linguistique du Centre et de l'Ouest - CerLICO

Référence électronique
Laurence Kister et Evelyne Jacquey, « NdeN et acquisition d'informations lexicales à partir du Trésor de la Langue Française Informatisé », *Corela* [En ligne], 5-1 | 2007, mis en ligne le 19 juin 2007, consulté le 30 avril 2019. URL : http://journals.openedition.org/corela/332 ; DOI : 10.4000/corela.332

NdeN et acquisition d'informations lexicales à partir du Trésor de la Langue Française Informatisé

Laurence Kister et Evelyne Jacquey

Introduction

1 Les systèmes automatiques d'extraction de l'information, d'indexation, de génération de thesauri, de résumé, etc. tendent à se multiplier. Pour ces systèmes, l'un des problèmes linguistiques majeurs est la reconnaissance des référents, c'est-à-dire des objets et des concepts désignés, au fil du texte, par des expressions linguistiques différentes et des mécanismes référentiels distincts (Amsili, Denis et Roussarie, 2005). La nécessité de résoudre cette question apparaît comme évidente pour repérer les thèmes de discours et pour qualifier plus précisément le contenu sémantique des documents. Parmi les différents mécanismes référentiels qui participent à la désignation des objets et des concepts, nous nous attachons à l'anaphore c'est-à-dire la relation qui s'établit entre "une expression dont l'interprétation référentielle dépend d'une autre expression (ou d'autres expressions) mentionnée dans le contexte et généralement appelée son antécédent" (Kleiber, 1994). Dans un texte, les relations successives qui s'établissent entre un antécédent et une ou plusieurs expressions anaphoriques forment une *chaîne de référence* (Boudreau et Kittredge, 2005). L'exemple ci-dessous illustre l'utilisation de plusieurs groupes nominaux et mots grammaticaux : *le schéma* est repris par *duquel*, *un élargissement du modèle guillaumien* et *il.*

> Ces problèmes sont traités par l'auteur dans un cadre théorique précis, issu de la psychomécanique du langage. Mais, [le schéma] à partir [duquel] cette étude multidirectionnelle s'organise est [un élargissement du modèle guillaumien] en ce sens qu'[il] intègre certains développements récents de la sémantique logique et les principaux acquis de la pragmatique. (Cervoni, 1991)

2 Plus généralement, l'étude des relations anaphoriques fait intervenir différents paramètres concernant l'expression anaphorique et le référent : (1) leur forme de surface,

(2) leur position syntaxique, (3) le ou les déterminants du référent qui conditionnent le mode de donation du référent (Kleiber, 1994) donc la saillance de celui-ci (Ariel, 1990) et (4) le contenu sémantique des constituants de l'expression référentielle. Nous nous intéressons ici à un type d'anaphore particulier où le référent est un groupe nominal de la forme *dét. N1 de (dét.) N2* (NdeN) et l'est un pronom sujet ([*L'affaire des bannis*] *prend tout son sens.* [*Elle*] *ravive de vieilles blessures*), un pronom objet ([*L'affaire* [*des bannis*]$_2$]$_1$ *prend tout son sens.* [*Elle*]$_1$ [*leur*]$_2$ *rappelle d'affreux souvenirs*), un pronom relatif (*Ils agissent dans la limite* [*des pouvoirs*] [*qui*] *leur sont conférés*), ou une structure démonstrative ([*La volonté*] *de la jeune fille s'oppose à* [*celle de*] *ses parents*). Pour cette forme particulière de référent, l'un des problèmes intéressants est de savoir si l'expression anaphorique coréfère au premier constituant (N1), au second (N2) ou à l'ensemble du groupe nominal NdeN.

3 La résolution de la coréférence entre les référents en NdeN et les expressions anaphoriques pronominales soulève plusieurs problèmes. S'il est relativement aisé de repérer automatiquement les pronoms d'un texte et d'en faire une analyse syntaxique, il reste délicat de reconnaître ceux qui sont des expressions anaphoriques. Si on admet ce problème résolu, on se trouve confronté à celui de l'identification de l'expression référentielle coréférente, à savoir la reprise du N1, du N2 ou du NdeN. L'examen des contenus sémantiques des expressions référentielles de la forme NdeN pose alors problème dans la mesure où nous ne disposons pas, actuellement, de lexique sémantique complet ni d'étiqueteur sémantique. Face à ce manque d'outil, notre préoccupation est d'établir des heuristiques solides en ce qui concerne la question du contenu sémantique des substantifs qui constituent les référents de la forme NdeN. Nous nous appuyons sur les heuristiques proposées lors de travaux personnels antérieurs pour trois autres critères : la forme de surface des expressions référentielles et des expressions anaphoriques, la position syntaxique des expressions référentielles et des expressions anaphoriques et les déterminants du NdeN. Nous montrons ensuite que le contenu sémantique des substantifs doit être déterminé du point de vue de l'opposition [+concret]/[+abstrait] et spécifions une procédure automatisable d'acquisition de la valeur de ce trait sémantique à partir d'un dictionnaire de langue informatisé, en l'occurrence le TLFi dans sa version XML catégorisée (définitions et exemples).

1. Résultats d'investigations antérieures

4 Les pistes envisagées au cours de travaux antérieurs émanent d'une analyse sur corpus[1] qui montre que le pronom relatif sujet *qui* est le plus productif (48,82 % des cas d'anaphores observées contre 27,79 % lorsque l'expression anaphorique est un pronom personnel sujet ou objet et 16,01 % pour un groupe nominal démonstratif). Cette répartition est à l'origine de notre choix de ne traiter que le cas des expressions anaphorique de la forme *qui*. En ce qui concerne les critères relatifs à l'expression référentielle (position grammaticale et détermination), dans 47,06 % des cas, l'expression anaphorique *qui* a pour antécédent une expression référentielle en position d'objet direct contre 14,94 % en position sujet, 14,38 % en position d'objet indirect et 23,62 % dans une autre position syntaxique. Pour les déterminants, l'article indéfini attire l'anaphorique (82 % de reprise du N2 pour les leNd'unN, 78 % de reprise de tout le groupe pour unNduN), l'absence de déterminant devant le N2 est un paramètre un faveur de la reprise de tout le groupe (76 % pour leNdeN et 91 % pour unNdeN). Le cas le moins tranché est

celui de l'article défini puisque 54 % des N2 sont repris quand le référent est de la forme leNduN.

5 Devant les caractères insuffisants des prédictions de saisie avec ces critères, même utilisés simultanément, nous nous intéressons aux contenus sémantiques. Des travaux sur le français et l'anglais ((Zagar, 1995), (Zagar et al., 1997) et (Mitchell et al., 1990)) en ce qui concerne la reconnaissance de l'antécédent du relatif *qui* dans des NdeN montrent que le caractère animé attire l'anaphore. Les premiers comptages que nous avons effectués sur corpus sont en accord avec ces observations. Cependant, il est nécessaire de les nuancer dans la mesure où les NdeN faisant l'objet d'une reprise par *qui* et dont au moins un des deux N est marqué [+animé] ne représentent que 9.2 % des occurrences (351 occurrences sur 1613) et que ceux qui contiennent au moins un N marqués [+inanimé] représentent 14.87 % des occurrences (240 sur 1613). Par ailleurs, nos observations font apparaître que 94.66 % de NdeN comportent au moins un N porteur du trait [+abstrait]. Nous proposons donc de ne pas opposer les traits [+animé] et [+inanimé] mais plutôt de les regrouper dans une catégorie plus générale [+concret]. L'opposition à considérer est alors [+abstrait] vs. [+concret] : celle-ci permet d'opposer ce qui est *matériel* c'est-à-dire ce qui peut être qualifié d'*objet* de la réalité à ce qui est *immatériel*. Une fois le regroupement effectué, nous obtenons 34,97 % occurrences (564 sur 1613) avec au moins un des N marqué [+concret] et 94.60 % des occurrences (1527 sur 1613) avec au moins un des N marqué [+abstrait]. Les NdeN composés de deux N [+concret] représentent 5.33 % (80 sur 1613) et ceux avec deux N [+abstrait] 65.03 % (1049 sur 1613). Cette approche nous conduit à considérer que quand un NdeN comporte un N [+concret] nous reprenons préférentiellement ce N.

2. Acquisition des traits sémantiques [+concret] vs. [+abstrait] à partir du TLFi

6 Dans la section précédente, nous avons vu que lorsqu'un référent de la forme NdeN contient un nom [+concret], c'est ce dernier qui est préférentiellement repris par l'expression anaphorique. Afin d'évaluer cette hypothèse au-delà des 1613 anaphores annotées manuellement dans des travaux antérieurs (*cf.* section 2), il serait intéressant de pouvoir annoter automatiquement des corpus plus importants. Cependant, comme le soulignent plusieurs travaux (Ide et Véronis, 1998 et Véronis, 2000), la désambiguïsation sémantique d'un mot en contexte, et donc l'annotation sémantique, en sont encore au stade de l'expérimentation. Dans notre cas, pour envisager l'annotation des substantifs des NdeN, deux approches sont possibles : (1) établir et implémenter des heuristiques applicables en corpus et/ou (2) disposer de listes de noms [+concret] et de noms [+abstrait]. Convaincus que ces deux approches ne sont pas antagonistes et peuvent au contraire coopérer, nous commençons par la seconde approche en établissant semi-automatiquement une liste de substantifs ayant au moins un emploi [+concret]. Pour éviter de construire un lexique trop étroitement lié au problème étudié, nous avons choisi d'utiliser une ressource couvrant une partie suffisante du lexique du français, jouissant d'une validité linguistique raisonnable et disposant d'une structuration des données permettant une exploitation automatique. Ces trois propriétés sont satisfaites par le Trésor de la Langue Française informatisé (TLFi) réalisé à l'ATILF (Dendien et Pierrel, 2002).

7 Une fois un sous-ensemble de substantifs reconnus comme ayant au moins un emploi [+concret], nous faisons l'hypothèse, certes trop brutale mais c'est une hypothèse de départ, que l'ensemble des substantifs du TLFi se divise en deux sous-lexiques : celui contenant les substantifs ayant un emploi [+concret], dit par la suite "le sous-lexique des concrets" et celui contenant les autres substantifs, supposés [+abstrait]. Munis du sous-lexique des concrets, il est possible d'annoter des corpus de syntagmes nominaux de la forme NdeN en associant le trait [+concret] à toute instance d'un substantif figurant dans le sous-lexique des concrets et [+abstrait] à toute instance d'un substantif ne faisant pas partie de ce sous-lexique.

8 Afin d'évaluer les résultats de cette première expérience, nous comparons la liste des concrets obtenue à partir du TLFi avec les 1613 NdeN annotés dans des travaux antérieurs.

2.1. Le dictionnaire de langue, un objet imparfait mais exploitable

9 Selon Véronis, les dictionnaires de langue ne permettent pas de désambiguïser les mots en contexte (Véronis, 2001). Dans notre cas, ils ne permettent pas en l'état de déterminer si les substantifs des NdeN sont en emploi [+concret] ou [+abstrait]. Cependant, les dictionnaires de langue constituent une source de connaissances sur la langue non négligeable. Un dictionnaire comme le TLFi jouit d'une couverture remarquable tant du point de vue du nombre de mots du français décrits (près de 100 000) que de celui de la quantité et de la diversité des informations apportées (en particulier l'illustration de toute définition par un ou plusieurs exemples extraits de FRANTEXT). De plus, bien que chaque article reflète la culture de son rédacteur, on peut espérer que sur l'ensemble du dictionnaire, le nombre et la diversité des rédacteurs atténuent le caractère subjectif du contenu des articles. Enfin, l'approche proposée par Véronis (Véronis, 2004), qui se base essentiellement sur des corpus, ne constitue pas non plus la solution étant donné notre objectif. Dans son étude du nom *barrage*, il dispose d'au moins un point d'entrée dans les corpus qui est le nom *barrage* lui-même. Dans notre cas, le point d'entrée devrait être les traits [+concret] vs. [+abstrait]. Or, si nous adoptions cette méthode, nous n'aurions l'information relative aux traits [+abstrait] et [+concret] qu'après avoir annoté les corpus. C'est pourquoi, nous initions la procédure d'étiquetage sémantique sur un dictionnaire, ce qui n'exclut pas, une fois une liste de noms [+concret] et/ou [+abstrait] établie, d'analyser ensuite en corpus les contextes d'apparition de ces noms afin de définir des heuristiques permettant une meilleure annotation. Cela s'avère indispensable lorsqu'un même nom est susceptible d'apparaître avec une signification concrète et une signification abstraite.

2.2. Procédure d'acquisition d'un sous-lexique de substantifs [+concret]

10 La procédure d'acquisition expliquée dans cet article s'appuie sur plusieurs hypothèses :

- toute définition d'un article est supposée refléter un sens particulier du mot décrit,
- l'ensemble des définitions d'un article est vu comme un corpus à l'intérieur duquel nous supposons qu'il existe une cohérence suffisante pour pouvoir dégager des critères distinctifs du point de vue de l'opposition [+concret]/[+abstrait],

- nous faisons l'hypothèse de départ que les définitions de substantifs qui contiennent le mot *objet* reflètent un sens [+concret]. Dans cette approche, le mot *objet* est vu comme un *descripteur.*

11 La procédure elle-même est itérée jusqu'à ce que plus aucun nouveau descripteur correspondant à un emploi [+concret] ne soit stocké. A chaque itération, les entrées de la procédure sont :

- l'ensemble des articles de substantifs du TLFi (54000 lemmes environ),
- la liste des mots jouant le rôle de descripteurs qui sont à rechercher dans les textes de définitions[2] des articles de substantifs (cette liste est initiée par le lemme *objet* et s'enrichit au fur et à mesure des itérations successives),
- pour chaque descripteur, la distance optimale à laquelle il reste intéressant de lancer la recherche[3], mais quel que soit le descripteur, une distance qui n'excède jamais la troisième position, distance à partir de laquelle le bruit est trop important.

Les sorties de cette procédure sont de deux formes :

- la liste de substantifs atteints par le biais de la requête cherchant au moins un des descripteurs de la liste donnée en entrée, à la bonne distance et en éliminant les définitions de substantifs qui contiendraient l'un des mots exclus,
- lorsque la distance le permet (descripteur en deuxième ou troisième position), la liste des mots à gauche de chaque descripteur de la liste en entrée. Ces mots sont considérés comme pouvant jouer le rôle de descripteur dans l'itération suivante. Cette liste est elle-aussi établie manuellement et en sélectionnant uniquement les substantifs (exclusion des quantifieurs comme *masse de*, des prédicats ambigus comme *construction*, etc).

La procédure se déroule ensuite de manière itérative :

- lancement de la requête en tenant compte des contraintes en entrée (liste de descripteurs, distances et listes de mots exclus associées à ceux-ci),
- analyse, nettoyage et croisement des résultats en éliminant les doublons,
- insertion de la liste des descripteurs en entrée,
- établissement d'une nouvelle liste de descripteurs en prévision de l'itération suivante.

12 Ci-dessous, nous proposons un échantillon du corpus exploité où les définitions concernent le substantif masculin *peigne*. Dans cet exemple, avec descripteur initial *objet*, seule la première définition serait repérée, donnant lieu au stockage d'une seule occurrence du substantif *peigne*.

> Subst. masc., **PEIGNE**, ***Objet*** de toilette ou de parure.
> Subst. masc., **PEIGNE**, Instrument de torture
> Subst. masc., **PEIGNE**, Pièce d'une machine agricole munie de fortes pointes, herse.
> Subst. masc., **PEIGNE**, Pointes des échalas qui dans un treillage dépassent les lattes horizontales
> Subst. masc., **PEIGNE**, Mécanisme formé d' un ensemble de lames dont chacune donne une note.
> Subst. masc., **PEIGNE**, Outil à fileter.
> Subst. masc., **PEIGNE**, Tringle en corne, en cuir, ou en acier qui est garnie de dents et que le peintre décorateur emploie pour faire le faux - bois en figurant les veines

13 Le second exemple ci-dessous montre d'une part comment les substantifs sont stockés : les entrées de substantifs *peigne, pique1* et *pouillerie* sont retenues car elles ont au moins une définition qui contient *objet* en première position avec *peigne*, en seconde avec *pique1* et en troisième avec *pouillerie*. En revanche, la définition de *patinoire* donnée dans cet exemple ne permet pas de stocker ce substantif dans le lexique des substantifs ayant au moins un emploi concret.

Subst. masc., **PEIGNE**, **Objet** de toilette ou de parure.
Subst. fém. **PIQUE1**, Tout **objet** terminé par une pointe métallique affectant la forme d' un fer de pique
Subst. fém. **POUILLERIE**, *Lieu* ou **objet** misérable , sordide .
Subst. fém., **PATINOIRE**, Lieu aménagé pour pratiquer le patinage sur glace

14 Cet exemple illustre aussi la détection et le stockage des descripteurs pour la prochaine itération. Le substantif *objet* se trouvant respectivement en seconde et en troisième position avec les définitions des substantifs *pique1* et *pouillerie*, la question se pose de savoir si les mots qui se trouvent avant *objet* sont des descripteurs. Les définitions étant étiquetées morphosyntaxiquement, nous sommes en mesure de rejeter *tout* et d'accepter *lieu* dans la mesure où celui-ci est étiqueté comme substantif. La fin des itérations est déclenchée dès que plus aucun nouveau descripteur n'est stocké.

2.3. Résultats obtenus

15 La première expérience utilisant cette procédure d'acquisition a permis d'aboutir à un résultat encourageant, mais pas suffisant. Le premier élément positif de cette expérience est la convergence des résultats, en relativement peu d'itérations (13 en tout). Cependant, ces résultats se présentent sous la forme d'un lexique extrêmement bruité et trop pauvre en informations. Des substantifs comme *intention* atteints par des définitions de la forme *objet de [+abstrait]*, telles que *Objet de pensée auquel l'esprit s'applique*, sont stockés aussi bien que des substantifs comme *éventail* atteints par des définitions de la forme *objet de matière légère [...]*.

Dans l'expérience suivante, nous introduisons deux étapes supplémentaires :

le contrôle des descripteurs à rechercher en cours de procédure afin d'éliminer ceux qui sont clairement ambigus entre [+abstrait] et [+concret] et afin d'éliminer les définitions de la forme *objet de Subst[+abstrait],*
le stockage de la définition et de l'ensemble de son contexte local afin de pouvoir vérifier la précision de l'acquisition a posteriori.

16 Pour évaluer la concrétude des futurs descripteurs à rechercher, nous nous appuyons sur la synthèse des critères utilisables proposés dans la littérature (Kister et Jacquey, 2006). A l'issue de cette seconde expérience, nous obtenons un lexique d'environ 14.000 substantifs (soit ≈ 45 % du total des substantifs du TLF) détectés par le biais de 27.400 définitions (soit ≈ 29 % du total des définitions de substantifs du TLF). En contrôlant l'acquisition des futurs descripteurs à rechercher, nous en avons sélectionné 514 et exclus 300, dont certains sont clairement [+abstrait] comme *exigence* ou bien ambigus entre [+abstrait] et [+concret] comme *malformation* ([+processus] ou [+résultat]).

17 La comparaison entre l'étiquetage fourni par ce lexique de substantifs ayant au moins un emploi concret et celui de notre corpus de référence permet d'évaluer la précision du lexique obtenu par la procédure d'acquisition décrite dans la section précédente : 62 % des substantifs sont correctement étiquetés. Cette comparaison permet de qualifier trois grands types d'erreurs qui correspondent à du silence (c'est-à-dire des substantifs, ou certains de leurs emplois, absents du lexique des concrets acquis automatiquement) : 12 % des erreurs sont dues à l'ambiguïté des mots vedettes entre [+concret] et [+abstrait] et à notre choix de rejeter les indices lexicaux ambigus, 8 % sont dues à une ambiguïté catégorielle de la vedette et 11 % sont dues à la présence d'un indice lexical indétectable dans les définitions des emplois concrets des mots vedettes. L'ambiguïté [+concret]/

[+abstrait] correspond en grande majorité à deux classes bien connues d'ambiguïtés lexicales sémantiques :

> l'ambiguïté [+processus]/[+résultat] avec des vedettes comme construction, la construction de la plate-forme vs. la construction quadrangulaire,
> l'ambiguïté [+objet]/[+collectif humain] avec des vedettes comme direction, la direction du véhicule vs. la direction du journal.

Concernant les indices lexicaux d'emplois concrets indétectables, deux cas se présentent principalement :

> des erreurs d'étiquetage apparaissent de manière régulièrement dans le TLFi XML catégorisé, c'est par exemple le cas de *personne* toujours étiqueté comme préposition lorsqu'il apparaît en début de définition, étiquetage qui interdit donc la sélection de toutes les définitions commençant par *personne qui [...]*
> des erreurs dues à l'absence de l'indice lexical qu'on aurait aimé stocker avec au moins l'un des 514 indices lexicaux sélectionnés au cours de la seconde expérience, c'est le cas de plusieurs indices intéressants comme *bâtiment, liquide*, etc.

18 Si on corrige les erreurs ainsi qualifiées, excepté celles qui concernent l'ambiguïté lexicale sémantique de certains mots vedettes (12 % des erreurs), on atteint un degré de précision de l'étiquetage de 81 %.

3. Spécifications pour une procédure d'acquisition plus précise

19 Un aspect important concerne la qualification du type de résultat que l'on peut s'attendre à obtenir en fonction de la distance des descripteurs. Jusqu'à une distance de 2 (le descripteur est en première ou seconde position dans le texte de définition), les substantifs obtenus sont majoritairement des hyponymes du descripteur. Lorsque la distance est de 2, deux catégories sont majoritaires en première position : les déterminants et les adjectifs qualificatifs. A partir d'une distance de 3, les possibilités se diversifient. On trouve parmi d'autres :

> des synonymes ou quasi-synonymes au sein de structures énumératives ou alternatives, c'est le cas de la définition *Partenaire ou adversaire avec lequel on engage une discussion [...]* dans l'article du substantif *interlocuteur* et avec le descripteur *adversaire*,
> des quantificateurs comme *masse de* avec la définition *masse de fer aciérée supportée par un billot* dans l'article du substantif *enclume* et avec le descripteur *fer*,
> des prédicats pour lesquels le descripteur est un actant comme *jeu* dans la définition *Petit jeu d'enfant, distraction puérile* dans l'article de *amusette* avec le descripteur *enfant*.

20 Une liste exhaustive de ces cas de figure demande l'application de techniques classiques d'analyse de corpus : étiquetage morphosyntaxique, alignement et intégration des nouveaux critères dans la procédure d'acquisition. Enfin, une étude linguistique mais utilisant des techniques d'analyse de corpus doit permettre de classer les descripteurs à chaque itération en fonction de leur degré de généralité et de leur productivité dans un premier temps.

Conclusion et perspectives

21 Cet article s'inscrit dans la problématique générale de l'étiquetage sémantique de corpus et de l'acquisition automatisée de ressources lexicales sémantiques à partir de ressources

lexicographiques structurées. La volonté d'étiqueter sémantiquement les corpus est dans notre cas justifiée par la contribution apportée pour l'évaluation des hypothèses courantes concernant la résolution des anaphores pronominales sur des antécédents de la forme *Det1 N1 de (Det2) N2* (NdeN). En l'occurrence, l'hypothèse à tester était dans cet article la préférence de l'anaphore pour le substantif [+concret] dans un syntagme de la forme NdeN. Afin de pouvoir dépasser l'annotation manuelle, il est nécessaire d'établir des procédures d'annotation sémantique. A cette fin, l'approche suivie est d'évaluer la faisabilité de l'acquisition automatique d'une liste de substantifs concrets à partir de l'ensemble des substantifs du TLFi.

22 L'expérience a montré que s'appuyer uniquement sur l'information fournie par les objets textuels du TLFi, en l'occurrence les objets "code grammatical" et "texte de définition", aboutit à un lexique caractérisé par un bon degré de couverture par rapport au corpus de référence (1613 occurrences de NdeN annotés manuellement, dont 156 noms annotés [+concret]). Cette expérience a cependant mis en lumière plusieurs lacunes : le coût humain, la progression inquiétante de la taille du lexique et celle des listes de descripteurs, la structure totalement plate du lexique obtenu et plusieurs erreurs dues à la non prise en compte d'éléments spécifiques du TLFi.

23 Cette expérience avait aussi pour but d'évaluer la qualité d'une approche très facile à automatiser complètement. Elle a donc permis d'établir plus précisément les spécifications d'une seconde procédure qui tirerait profit, non seulement de la structure des données lexicographiques mais aussi, appliquerait des traitements de corpus afin de mieux qualifier les descripteurs et leurs contextes d'apparition.

24 Enfin, cette étude pourra être étendue à l'examen des noms abstraits. On rappelle en effet que les syntagmes NdeN comportant un nom concret sont loin de refléter la configuration majoritaire pour les relations anaphoriques. Dans plus de la moitié des cas, les deux noms du NdeN sont abstraits et aucune hypothèse tranchée n'est faite actuellement avec deux noms abstraits quand au choix de l'antécédent (soit le second nom, soit le premier, soit l'ensemble du syntagme complexe). Pour résoudre cette question, une piste à suivre peut être de considérer la qualification et la classification des substantifs abstraits, point sur lequel la procédure d'acquisition de sous-lexique telle que nous l'envisageons maintenant pourra apporter des éléments de réponses.

BIBLIOGRAPHIE

Amsili P., Denis P. zt Roussarie L. (2005), « Anaphores abstraites en français : représentation formelle », in Modèles et algorithmes pour la résolution d'anaphores, J. Busquets et d. Hardt (eds), Traitement automatique des langues, vol. 46/1, pp. 15-39.

Ariel M. (1990), Accessing Noun-phrase Antecedents, London, Roufledge.

Boudreau S., Kittredge R. (2005), « Résolution des anaphores et détermination des chaînes de coréférences : différences entre variétés de textes », in Modèles et algorithmes pour la résolution

d'anaphores, J. Busquets et D.Hardt (eds), Traitement automatique des langues, vol. 46/1, pp. 41-69.

Cervoni J. (1991), La Préposition : étude sémantique et pragmatique, Louvain-la-Neuve, Editions Duculot.

Dendien J., Pierrel J-M. (2002), « Le trésor de la langue informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence », TAL.

Ide N., Véronis J. (1998), « Word Sense Desambiguisation: The State of Art », Special Issue of Computational Linguistics, vol. 24(1), pp. 1-40.

Kister L., Jacquey E. (2006), Traits sémantiques et anaphores pronominales, 4ème Rencontres de sémantique et de pragmatique, Orléans, 13-15 juin 2006.

Kleiber G. (1994), Anaphores et pronoms, Champs linguistique, Louvain-la-Neuve, Editions Duculot.

Mitchell D.C., Cuetos F. et Zagar D. (1990), « Reading in Different Languages: is there a Universal Mechamism for Parsing Sentences », Comprehension Processus in Reading , D.A. Balota, G.B. Flores d'Arcais and K. Rayner (eds.), Hillsdale, Lawrence Erlbaum, Associates, Inc.

Véronis J. (2000), « Annotation automatique de corpus : panorama et état de la technique », Ingénieurie des langues, Pierrel J-M (ed), Paris, Hermès.

Véronis J. (2001), « Sense tagging: does it make sens? », Actes de Corpus Linguistics'2001.

Véronis J. (2004), « Quels dictionnaires pour l'étiquetage sémantique », Le français moderne, 2004/1, pp. 27-38.

Zagar D. (1995), « La lecture : processus de base », Habilitation à diriger les recherches, U.F.R. de Sciences Humaines - Université de Bourgogne, LEAD - CNRS Ura 1938.

Zagar D., Pynte J. et Rativeau S. (1997), « Evidence for Early-closure Attachment on First-pas Reading Times in French » , The Quaterly Journal of Expérimental Psychology, 2, p. 421-438.

NOTES

1. Le corpus comporte 5365 relations anaphoriques dont 1613 réalisées au moyen du pronom relatif *qui*. Il se compose de textes scientifiques et techniques, de mémoires, d'articles de revues spécialisées, de rapports d'activité, de romans, d'articles (presse quotidienne, presse interne d'entreprise, magazines de large diffusion, journaux d'information destinés à des clients ou adhérents) et de dépêches relatant des catastrophes naturelles.

2. L'objet textuel "texte de définition" a été privilégié par rapport à son objet englobant "définition" car les contenus textuels de l'objet "définition" peuvent commencer par deux virgules ou des guillemets qui sont comptés comme deux mots, ce qui fausse ensuite la recherche. Au contraire, les objets textuels "texte de définition" commencent toujours par un lexème (nom, adjectif, déterminant, verbe, préposition, adverbe, etc).

3. Chaque descripteur ne peut en effet pas être recherché avec la même efficacité dans toutes les positions d'un texte de définition. Une recherche du mot *objet* au-delà de la troisième position par exemple produit trop de bruit (ex: Agriculture Activité$_0$ ayant$_1$ pour$_2$ objet$_3$: [...]).

RÉSUMÉS

Pour les systèmes de TAL, traiter la résolution de la référence s'impose pour repérer les thèmes qui qualifient le contenu sémantique des documents. Dans cet article, nous proposons une manière d'acquérir des informations sémantiques pour résoudre les anaphores où l'expression référentielle est de la forme NdeN et l'expression anaphorique est un pronom relatif sujet. Afin de prendre en compte le contenu sémantique des noms des NdeN, nous spécifions une méthode automatisable de construction d'une liste de noms concrets du français à partir des définitions du TLFi dans sa version XML catégorisée. Ce sous-lexique permet ensuite d'étiqueter sémantiquement les corpus et de prédire le référent du pronom relatif sujet (35% des cas font intervenir un nom concret).

For NLP systems, a major issue consists in resolving reference in order to find the themes of documents. In this article, we present a way to find semantic informations to resolve anaphors which use a referential expression of the form NdeN and an anaphor realized by a subject relative pronoun. This method is based on the semantic content of nouns in NdeN groups and uses a list of concrete nouns in French which can be automatically extracted from the definitions of the TLFi dictionary in its XML-tagged version. Such a list is then used to annotate corpora in order to predict the selection of the good nominal referent which can be the second noun or the entire NP (35% of referential expressions contain a concrete noun).

INDEX

Keywords : anaphor, NdeN, semantic tagging, lexicon extraction, lexical semantics
Mots-clés : anaphore, NdeN, étiquetage sémantique, extraction de lexique, sémantique lexicale

AUTEURS

LAURENCE KISTER

Nancy Université – Atilf UMR 7118

EVELYNE JACQUEY

Nancy Université – Atilf UMR 7118